DOCUMENT RESUME

ED 137 319 95 TM 005 959

AUTHOR Hannan, Michael T.; And Others
TITLE Specification Bias Analysis of the Effects of Grouping of Observations in Multiple Regression Models.
INSTITUTION Vasquez Associates Ltd., Milwaukee, Wis.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Apr 75
CONTRACT NIE-C-74-0123
NOTE 33p.; Paper presented at the Annual Meeting of the American Educational Research Association (59th, Washington, D.C., April 1975)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Multiple Regression Analysis; *Research Methodology; Sampling; *Statistical Bias
IDENTIFIERS *Grouping (Statistical)

ABSTRACT

Grouping is a statistical procedure through which members of the same group are considered as a single unit of observation. There are various ways to assign group membership and various ways to assign values of variables to groups. There are methodological problems associated with grouping in general and with particular methods of grouping. This paper argues that a wide variety of complex analytical problems concerning inferences from grouped observations can be understood from the use of a few simple principles. The paper focuses on multiple regression models which use grouping and shows that the effects of grouping depend centrally on the quality of the specification of the regression model used. Simulated examples as well as examples from the literature are presented and discussed. The principles developed are then extended to more complex cases. In particular, estimation from grouped observations in systems of simultaneous equations and in dynamic models for panel analysis are examined. (Author/JKS)

SP

Specification Bias Analysis of the Effects of Grouping of Observations in Multiple Regression Models*

Michael T. Hannan
Alice A. Young
Francois Nielsen

Stanford University

Technical Report No. 2

April 1975

Consortium on Methodology for Aggregating Data in Educational Research

*This research was conducted under National Institute of Education contract #NIE-C-74-0123.

Vasquez Associates, Ltd.
1744 N. Farwell Ave.
Milwaukee, Wisconsin 53202

Specification Bias Analysis of the Effects of Grouping of Observations in Multiple Regression Models*

Michael T. Hannan
Alice A. Young
Francois Nielsen
Stanford University

Paper Presented at the Meetings of the
American Educational Research Association,
Washington, D.C., April, 1975

*The research reported here was conducted under National Institute of Education contract NIE-C-75-0123. Bonnie Ecker assisted us in the computations reported.

The difficulties involved in making inferences across units of analysis have been discussed in every social science (Hannan, 1971). Despite the considerable methodological literature that has developed, research practice appears substantially unchanged. Research on educational organization and the consequences of education is by no means an exception (Burstein, 1975). So it may be useful to continue to restate the difficulties involved in making inferences across analysis. We focus on those aspects of the general problem that arise in estimation from grouped observations.

The point of the paper is to argue that a wide variety of complex analytic problems concerning inferences from grouped observations can be understood by use of a few simple principles. To make this argument, we restate the available results in simple terms. The thrust of the earlier work is to show that the effects of grouping depend centrally on the quality of model specification. To reinforce this perspective, we present the results of a Monte Carlo simulation and analysis of empirical data. Then we show that the simple principles can be extended in a straightfoward manner to the analysis of more complex cases than have been addressed in the existing literature. We treat two cases: estimation from grouped observations in systems of simultaneous equations and in dynamic models for panel analysis.

II. Results From the Two Variable Regression Model

We will first briefly restate the well known results for the two variable regression model:

$$Y = \alpha + \beta X + u \tag{1}$$

where the disturbance, u, has mean zero, constant variance and is asymptotically uncorrelated with the regressor X. The least squares estimator

$$b = \frac{\Sigma (X_{ij} - X_{i.}) Y_{ij}}{\Sigma (X_{ij} - X_{i.})^2}$$

is a consistent estimator for β of (1).

Consider the grouped regression for:[1]

$$\bar{Y} = \alpha + \beta \bar{X} + \bar{u}$$

and the least squares estimator

$$\bar{b} = \frac{\Sigma (\bar{X}_j - \bar{X}) \bar{Y}_j}{\Sigma (\bar{X}_j - \bar{X})^2} \tag{2}$$

The effects of grouping of observations are to be evaluated by comparing properties of the two least squares estimators, b and $\bar{b}$. To make such comparisons we need to further specify the nature of the grouping process. Any method of grouping observations that retains the absence of correlation of the regressor and the disturbance, i.e. between $\bar{X}$ and $\bar{u}$, will yield consistent estimation of β.

Three cases deserve mention: random grouping, grouping that maximizes variation in X (grouping "by X") and grouping that maximizes variation in Y (grouping "by Y"). It is widely noted (Prais and Aitchison, 1954; Blalock, 1964; Hannan, 1971; Feige and Watts, 1972) that both random grouping and grouping by X will preserve the lack of correlation between $\bar{X}$ and $\bar{u}$ and as a result for both methods, plim $\bar{b}$ = plim b = β.

Blalock (1964) was apparantly the first analyst to point out that grouping by Y will tend to produce a correlation between $\bar{X}$ and $\bar{u}$ even when X and u are uncorrelated (in the sample).

[1] Throughout we assume equal sized groups. For a treatment of efficient estimation with unequal sized groups, see Prais and Aitchison (1954).

This means that least squares applied to (2) will be inconsistent in this case (plim $\bar{b} \neq b = \beta$).

The results on efficiency of least squares estimators applied to (2) are also well known (Cramer, 1964; Hannan and Burstein, 1973; Feige and Watts, 1972). Random grouping eliminates systematic as well as error variation indiscriminately and is as a result considerably more damaging to efficiency of L.S. than is grouping by X. In fact, grouping by X is optional in the sense that no other grouping method can yield a grouped L.S. estimator with smaller variance.

III. Grouping and Specification Bias

Published (and unpublished) results on grouped and ungrouped data typically show considerable divergence. In many instances this is the case even when it seems unlikely that the data were grouped systematically by values of the dependent variable. Some more general phenomenon seems to be involved. We argued earlier (Hannan, 1971; Hannan and Burstein, 1973) that grouping may give rise to systematic bias when the specification of the ungrouped model is faulty. In fact, we argue, grouping may tend to magnify errors of specification.

To show this, we consider a simple extension of the model considered earlier:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + u \qquad (3)$$

where we assume that the disturbance is asymptatically uncorrelated with each regressor, X_1 and X_2. Suppose that a researcher fails to include X_2 and instead estimates a model

$$Y = \alpha' + \beta_1' X_1 + w. \qquad (4)$$

Following Theil (1957) we find:

$$\text{plim } b_1' = \beta_1 + \beta_2 b_{21} \qquad (5)$$

where b_{21} is the sample regression of X_2 and X_1, (the coefficient of what Theil calls the auxiliary regression). As long as the two regressors are correlated, least squares applied to (4) will give inconsistent estimates of β_1 in (3). The magnitude of the discrepancy

$$\text{plim } b_1' - \beta_1 = \beta_2 b_{21}$$

is called the specification bias of the estimator (as an estimator of β_1).

The grouped analogues to (3) and (4) are

$$\bar{Y} = \alpha + \beta_1 \bar{X}_1 + \beta_2 \bar{X}_2 + \bar{v} \qquad (6)$$

and

$$Y = \alpha' + \beta_1' \bar{X}_1 + \bar{w} \qquad (7)$$

The properties of least squares applied to (6) can be understood by using results of the previous section. It is still the case that random grouping preserves unbiasedness of least squares but sacrifices efficiency, grouping by Y biases least squares, and grouping by one or more independent variable preserves unbiasedness and is more efficient than random grouping.[2] (The results of the simulation presented in Section III are relevant to this discussion)

Least squares applied to (7), yields an estimator $\bar{b}_1'$ such that

$$\text{plim } \bar{b}_1' = \beta_1 + \beta_2 \bar{b}_{21} \qquad (8)$$

As before, if $\bar{X}_1$ and $\bar{X}_2$ are correlated, least squares applied to (7) gives unconsistent estimation of β_1.

Next, contrast the two (inconsistent) estimators, b' and $\bar{b}'$. We see that,

$$\text{plim } (\bar{b}' - b') = \text{plim } (\beta_2[\bar{b}_{21} - b_{21}]). \qquad (9)$$

And note that the grouped and ungrouped results will diverge if $\bar{b}_{21}$ differs from b_{21}. But both $\bar{b}_{21}$ and b_{21} are coefficients from two variable regressions so we can use results from the previous section.[3] With both random grouping and grouping by X_1, plim $\bar{b}_{21}$ = plim b_{21}. As a result, plim $(\bar{b}'-b') = 0$ and on the average one will obtain the same (incorrect) results from the grouped and the ungrouped regressions.

---

[2] These results can be shown most easily by adopting the grouping matrix (transformation) introduced by Prais and Aitchison (1954) -- see also Feige and Watts (1972) The key is to analyze whether or not the transformation is systematically related to the disturbance. For example, when the independent variables are uncorrelated with the disturbance, grouping that maximizes variation in one or more of them will be unrelated to the disturbance.

[3] We will not contine to repeat that grouping by Y will always lead to inconsistent estimation.

Quite the contrary is the case for grouping by the omitted variable, X_2. For when observations are grouped systematically by X_2, this is grouping by the dependent variable from the perspective of the auxiliary regression. We noted in the previous section, grouping by the dependent variable leads to inconsistent estimation, i.e. plim $\bar{b}_1' \neq$ plim b_1'. As a result, the grouped and ungrouped least squares estimators differ in the limit and one will tend to draw different inferences from each. We conclude, then, that grouping by an omitted variable has consequences quite different from random grouping or grouping by an independent variable.

It is important to know whether there is a possibility that the grouped estimator would be closer (asymptotically) to β_1 than is the ungrouped estimator. For such an "aggregation gain" to occur, $|\bar{b}_{21}| < |b_{21}|$ (compare (5) and (6)). None of the cases we considered yield this result. Hannan and Burstein (1973) show that, when observations are grouped by rules that relate additively to variables in the model, grouping will magnify any specification bias in the ungrouped estimator. Nevertheless, the possibility exists that the data are grouped by some non-additive combination of variables (e.g. place into the same group those observations that are highest on X_1 but lowest on X_2) -- see the discussion in Hanushek et al. (1974) While such outcomes seem highly unlikely in "natural" groupings, we cannot conclude that grouping always magnifies specification bias.

It may seem unusual to compare the properties of alternative estimators of misspecified models. When we consider properties of estimators we routinely assume proper model specification. However, as researchers

we acknowledge the practical difficulties involved in arriving at the correct specification and treat substantive models as partial and tentative. In a sense, then, we commonly admit to likely misspecification of our models. In this light it seems important to realize that specification errors that may be small and not very damaging to inference in an ungrouped analysis may be magnified by grouping into very sizable errors.

In the next two sections we illustrate, first with a Monte Carlo simulation and then with analysis of real data, the considerable damage to inference that is possible.

IV. A Monte Carlo Simulation

To illustrate each of the points made in the proceeding sections we designed the following Monte Carlo simulation. We generated 100 samples of size 500 from a population characterized by

$$Y = 1X_1 + 1X_2 + \delta u$$

where X_1 and X_2 were both N(0,1), and u was N(0,1) distributed independently of X_1 and X_2, and δ was a constant set at two different levels to vary R^2 for the equation. We varied R^2 (at .3 and .7) and r_{12} (at .25, .50, and .75). For each of the six parameter combinations of R^2 and r_{12}, we then grouped observations into 50 equal sized groups in each sample generated by three methods: randomly, by values of X_1, and by values of X_2.[4]

We estimated (with ordinary least squares) each of the following regressions

$$Y = \beta_1 X_1 + \beta_2 X_2 + u \qquad (10)$$

$$Y = b_1 X_1 + v \qquad (11)$$

from ungrouped and grouped data (for each grouping method).

Before treating the misspecified model, consider first the results for the correctly specified equation, (10). As far as we know, no one has shown the consequences of grouping by one regressor in a multiple regression. As we see in Tables 1 and 3, grouping by X_1 or X_2 yields a pattern of estimates that centers relatively close to the true values. These results further confirm our claim that, except for grouping by

[4] In grouping by X_1 we ordered observations in decreasing order by X_1 values and placed the first 10 observations into one group, the next 10 in the next group, etc. Then each of the grouped observations was replaced with the group mean.

Y, grouping affects consistency only by magnifying specification bias. When, as here, there is no specification bias, grouping observations by one of a set of regressors does not alter the average value of the least squares estimators.

Earlier we noted that grouping by the regressor in a two variable regression is optimal in the sense that it minimizes variance of the estimator (among the class of consistent grouped estimators). In Table 4 we see that mean squared error (bias squared plus variance) of estimates over the 100 samples is considerably lower for grouping by either of the regressors, as contrasted with random grouping. Presumably grouping by both regressors simultaneously would further reduce the variance of the grouped estimator.

The results on mean error for the misspecified equation, Table 2, closely fit our expectations. Notice first the specification bias in the ungrouped estimator. Since both X_1 and X_2 are N(0,1), $b_{21} = r_{12}$ and plim $b_1' = 1 + 1r_{12}$ and what we are calling error is simply r_{12}. The data conform closely to this result. Grouping by X_1 gives mean errors almost identical to those for the ungrouped case. As we suggested earlier, grouping by X_1 in this case gives optimal estimates of the wrong term, $\beta_1 + \beta_2 b_{21}$. Finally, grouping by the omitted variable, X_2, greatly magnifies the specification bias. The magnitude of the inflation varies from a two-thirds increase (for $r_{12} = .75$) to a more than six-fold increase (for $r_{12} = .25$).[5]

[5] This pattern is not surprising since the lower r_{12}, the greater the upward bias in b_{21} from grouping by X_2 (as long as $r_{12}>0$), (cf. Blalock, 1964).

V. An Empirical Illustration

In a recent publication, Bidwell and Kassarda (1975) purport to demonstrate that organizational properties of school distinct (e.g. teacher/pupil ratio) affect student achievement by regressing district mean achievement on school district properties. They argue strenuously against including (aggregated) properties of individuals in these analysis:

> ...all measured relationships are at the level of the school district. None pertains to individuals, schools, or other sub-units of the district. The reader should keep in mind that we are not analyzing antecedents of the academic achievement of individual students, rather the overall effectiveness of a school district as measured by the aggregate achievement level for all its students at a given grade. Introducing multiple levels of analysis into the same model brings difficulties of estimation and interpretation (e.g. the "ecological fallacy") that we wish to avoid (Bidwell and Kassarda, 1975:63).

We have no quarrel with such organizational analysis. However, given the extensive knowledge available concerning the determinants of achievement at the individual level and the known correlation of at least some of them, e.g. SES, with school quality, we wonder if omitting all but school district properties from the model gives unrealistically large organizational effects. In fact, we argue that omitting other (correlated) causes of achievement together with estimation from the grouped observations produces exactly the type of fallacious inference Bidwell and Kassarda claim to avoid.

To see this, suppose that school district characteristics, SD, and SES background of students both affect their academic achievement, A:[6]

---

[6] Note that this is not Bidwell and Kassarda's model. But, if school district characteristics do affect achievement, they should appear in this sort of model. In a sense, Bidwell and Kassarda did introduce some input variables (though they consider them to be environmental constraints)--namely percent non-white and education of the adult population. They report that adult education (proportion of the population with at least four years of education had an insignificant effect in the achievement regression (although the beta-weights reported are sizable). Perhaps, the failure to measure education of parents accounts for the small effect of education. At any rate, Bidwell and Kassarda then excluded such inputs from their [illegible]

$$A = \alpha_0 + \alpha_1 SD + \alpha_2 SES + w$$

assuming that $r_{SD,SES} > 0$ and that w is distributed independently of the two regressors.

The models (i.e. relevant parts of the models) estimated by Bidwell and Kassarda take the form:

$$\bar{A} = \alpha_0' + \alpha_1' \bar{SD} + \bar{w}$$

(where by definition $\bar{SD} = SD$). To evaluate the properties of $\bar{a}_1'$, the least squares estimator of α_1' we refer to the results already presented. Elimination of SES from the ungrouped equation will clearly produce a specification bias in least squares estimation of the school district effect. What about magnification of this bias from grouping? If, as we suspect, individuals are selected into school districts on the basis of their SES, the observations in Bidwell and Kassarda's analysis were (to some extent) grouped by the omitted variable. If so, their reported school effects would greatly overstate the case (be upwardly biased).

We cannot, of course, determine from Bidwell and Kassarda's analysis how serious this problem is. We had access to data on sixth graders in California schools and school districts that enable us to illustrate the damage to inference. The outcome measure used is reading achievement (as in Bidwell and Kassarda's analysis). The structural variables available at the district level were resources (expenditures/average daily attendance), and pupil teacher ratios.[7] The input variables avail able were parents occupation (six-categories) and IQ estimates by teachers.

Our strategy was to construct two regression models: one regresses

[7] Following Bidwell and Kassarda we include percent non-white in the model as well.

achievement on only the school district structural properties (resources, pupil/teacher ratio) and percent non-white; the second adds the inputs variables. Then we estimate (with ordinary least squares) each regression with pupil level inputs and outputs, school mean inputs and outputs, and district mean inputs and outputs.[8] This allows us to contrast specification bias at each level.

The results are presented in Table 7. Consider first the improperly specified equation analogous to that used by Bidwell and Kassarda (the three columns on the left in Table 7). The beta-weights associated with district structural properties are enormously inflated with grouping. The increase from the pupil to the school level is more than 200% for each effect (from -.119 to -.465 and -.167 to -.368) and a second large increase from the school to district level (-.465 to -.664, and -.368 to -.737). Then, consider the effect of introducing the input variables. Their addition does not greatly alter the district property effects at the pupil level.[9] But, the effects of grouping are greatly reduced by the improvement in the specification. The change in the district structure effects is very much smaller, going from pupil to school level, (from -.149 to -.196 and from -.146 to -.048). So these results conform very closely to our expectations. They very clearly document the magnitude of the hazard to correct inference arising in estimation from grouped data and poorly specified models.

---

[8]The regression containing inputs cannot be estimated at the district level due to small number of observations.

[9]We suspect that adding more input variables would considerably reduce these effects, however.

VI Extensions

In this section we illustrate how the simple analytic results of the preceeding sections can be used to handle more complex models. We consider the two types of extensions most likely to be of interest to substantive researchers: simultaneous equations models and dynamic models.

A. A System of Simultaneous Equations

Causal systems in which there is reciprocal causation (e.g. aspirations affect achievement, and achievement affects aspirations) lead to complex inference problems. Let us confine the discussion to a simple case:

$$Y_1 = \alpha_{12}Y_2 + \alpha_{11}X_1 + w_1 \qquad (12)$$

$$Y_2 = \alpha_{21}Y_1 + \alpha_{22}X_2 + w_2 \qquad (13)$$

where w_1 and w_2 are distinguished independently of X_1 and X_2. It is easy to show (cf. Johnston, 1971) that the endogenous variables on the right hand side (Y_2 in (12), Y_1 in (13)) are correlated with the disturbances. Least squares applied to either (12) or (13) will be inconsistent--will contain "simultaneous equations bias."

Instead of using ordinary least squares, we solve for the so-called reduced-form:

$$Y_1 = \pi_{11}X_1 + \pi_{12}X_2 + q_1 \qquad (14)$$

$$Y_2 = \pi_{21}X_1 + \pi_{22}X_2 + q_2 \qquad (15)$$

ordinary least squares applied to (14) and (15) is consistent. Having obtained estimates of the reduced-form coefficients, we can, in this simple case, solve directly for the estimates of parameters of (12)-(13). More generally, we calculate $\hat{Y}_1$ and $\hat{Y}_2$ from estimated (14)-(15) and substi-

tute them for Y_1 and Y_2 on the RHS of (12)-(13) and then apply ordinary least squares to the revised system (12-13). The first method is called indirect least squares, the second, two-stage least squares. Both methods lead to consistent estimation.

Next consider the grouped data:

$$\bar{Y}_1 = \alpha_{12}\bar{Y}_2 + \alpha_{11}\bar{X}_. + \bar{w}_1 \qquad (16)$$

$$\bar{Y}_2 = \alpha_{21}\bar{Y}_1 + \alpha_{22}\bar{X}_2 + w_2 \qquad (17)$$

which has as its reduced-form,

$$\bar{Y}_1 = \pi_{11}\bar{X}_1 + \pi_{12}\bar{X}_2 + \bar{q}_1 \qquad (18)$$

$$\bar{Y}_2 = \pi_{21}\bar{X}_1 + \pi_{22}\bar{X}_2 + \bar{q}_2 \qquad (19)$$

If the data are grouped by values of X_1 or X_2, our earlier results imply that estimation of the grouped reduced-form will be asymptotically equivalent to estimation in the ungrouped reduced-form. Since estimates of the parameters of (16)-(17) are functions of the reduced-form estimates, both indirect least squares and two-stage least squares will be asymptotically equivalent for both grouped and ungrouped estimators.

Grouping by an endogenous variable, Y_1 or Y_2, is quite another matter. We have noted that grouping by a dependent variable leads to inconsistent estimation in that equation. So if we group by Y_1, say, the first equation in the reduced-form, (18), will be inconsistently estimated. Since the estimates of the coefficients of (16)-(17) are functions of <u>all</u> the reduced-form estimates, grouping by an endogenous variable will lead to inconsistent estimation throughout the structural form (16-17).

B. A Dynamic Model for Panel Analysis

Consider the following simple dynamic formulation[10] applied to panel observations on N individuals over T waves of observation (e.g. a cohort of students observed once a year for several years). Assuming that the causal structure is constant, we pool all NT observations into (20):[11]

$$Y_{it} = \alpha_0 Y_{i,t-k} + \alpha_1 X_{i,t-k} + \varepsilon_{it} \quad (i=1,..,N;\ t=1,..,T) \qquad (20)$$

It is usually the case that the disturbances in such models are positively correlated over time. If so, the disturbance will be correlated with at least one right hand side variable, Y_{t-k}, and least squares applied to (20) will be inconsistent. To go more deeply into the problem we must specify the nature of the process accounting for the autocorrelation. Assume that the disturbance has the following variance-components form (cf. Nerlove, 1971):

$$\varepsilon_{it} = \mu_i + v_{it} \qquad (21)$$

where μ_i is a time invariant constant associated with each unit (e.g. each pupil's unobserved characteristics) and uncorrelated with X_{t-k}, and v_{it} is a well behaved random N(0,1) disturbance uncorrelated with μ. This model has proven useful in panel analysis (cf. Nerlove, 1971; Hannan and Young, 1974).

It is helpful to restate the model in matrix form to highlight the parallel with the foregoing. Let

---

[10] The model is dynamic due to the presence of the lagged dependent variable --it is a stochastic difference equation.

[11] We pool observations to allow correction for the autocorrelation problem discussed below. The problem of autocorrelation remains whether or not one pools observations.

$$\underline{Z}_{t-k} = (\underline{Y}_{t-k}\ \underline{X}_{t-k})$$

where $\underline{Y}_{t-k}$ and $\underline{Y}_{t-k}$ are NX1 vectors

$$\underline{\alpha} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} ,$$

$$\mu = \begin{pmatrix} \mu_1 \\ \cdot \\ \cdot \\ \mu_N \end{pmatrix}$$

where the μ_i's are the coefficients corresponding to the dummy variables, contained in $\underline{\Delta} = (\delta_1, \delta_2, .., \delta_N)$, where δ_i is an Nx1 dummy variable for the ith individual.

Then, the model (20-21) can be written:

$$\underline{Y}_t = \underline{Z}_{t-k}\alpha' + \underline{\Delta}\,\mu + \underline{v}_t \qquad (22)$$

The usual least squares procedure employs only

$$\underline{Y}_t = \underline{Z}_{t-k}\alpha' + v'_t \qquad (23)$$

The omission of the set of dummy variables, $\underline{\mu}$, leads to the type of specification bias considered above.

Consider next the grouped versions of (22) and (23):[12]

$$\bar{\underline{Y}} = \bar{\underline{Z}}_{t-k}\,\underline{\alpha} + \bar{\underline{\Delta}}\,\underline{\mu} + \bar{\underline{v}}_t \qquad (24)$$

and

$$\bar{Y}_t = \bar{\underline{Z}}_{t-k}\alpha' + \bar{v}'_t \qquad (25)$$

Obviously least squares applied to (25), a misspecified model, will be inconsistent. But will the results differ from those of the ungrouped case? As always, it depends on the nature of the grouping process. Our previous results follow immediately from a generalization of Theil's

[12] For a treatment of (24) in terms of grouping matrices, see Appendix A.

(1957) specification analysis. If the data are grouped by Z_{t-k} (either X or Y) there is no magnification of the specification bias from the ungrouped estimator. If the data are grouped by μ (i.e. if the unobserved properties of the units are related to the selection of the group), the bias from the ungrouped estimator is magnified.

Finally, note that least squares applied to (22) is a consistent estimator as is least squares applied to (24) -- we call these least squares with dummy variables (LSDV). These estimators are asymptotically equivalent to generalized least squares estimators (Ameniya, 1967; see also Hannan and Young, 1974).

Conclusion

Comparisons of grouped and ungrouped estimators for models that are incorrectly specified clarifies the manner in which grouping affects inferences from regression analysis. In particular, under quite general conditions when observations are grouped by some criterion that miximizes variation in an omitted variable (correlated with included regressors), grouping magnifies the specification bias of the ungrouped estimation. Both our Monte Carlo simulation and substantive analysis suggest that the magnification can be quite large. As a result qualitative inferences from regressions with grouped data may differ greatly from those made from regressions with ungrouped data.

The principles used in the specification bias analysis of the effects of grouping can be used to understand the effects of grouping in more complex models. In particular, we have shown that grouping in a simultaneous equations model and in a dynamic model for panel analysis introduces no complications that cannot be addressed with this methodology.

Appendix A

Estimation with Grouped Data in a Pooled Cross-Sections and Time Series Model.

The effect of aggregation in the context of the pooling of cross-sections may be conveniently analyzed with the notations introduced by Prais & Aitchison (1954). The model we shall examine is the one defined in Chapter VI, B. In matrix notation:

$$\underline{Y}_t = \underline{Z}_{t-k}\ \alpha + \underline{\Delta}\ \mu + \underline{V}_t \qquad \text{(A1)}$$

where $\underline{\Delta} = (\delta_1, \delta_2, \ldots, \delta_N)$

If the data are arranged so that the N rows corresponding to individual observations at the first time period are placed first, the N rows corresponding to the second time period next, etc..., these $\underline{\Delta}$ may be written explicitly in matrix notation as:

$$\underline{\Delta} = 1_T \otimes I_N \qquad \text{(A2)}$$

1_T is a T x 1 vector of 1's.

I_N is the N x X identity matrix

$\otimes$ denotes the Kronecker product or direct sum.

The aggregation rule which one is most likely to use in this context consists of taking overages over the same individuals at each time period (for example, the academic performances of students measured over time are aggregated at the classroom level). In that case, and if one assumes that the number of individual observations in each group is the same, the aggregation procedure may be represented by the linear transformation:

$$M = I_m \otimes 1/n \cdot 1_n' \qquad \text{(A3)}$$

applied to the data corresponding to each time period.

m is the number of groups

n is the number of individuals within a group. ($n \cdot m = N$)

Since there are T time periods, the complete transformation becomes:

$$G = I_T \otimes M = I_T \otimes (I_m \otimes 1/n \cdot 1_n')$$

The grouped model may now be written as:

$$G.\underline{Y}_t = G\,\underline{Z}_{t-k} \cdot \alpha + G\,\underline{\Delta}\,\underline{\mu} + \underline{V}_t \qquad (A4)$$

It is important to investigate the effect of aggregation on the error-components aspects of the pooled model, particularly the extent to which the specific methods developed by Nerlove (1971) and Hannan & Young (1974) have to be modified. To do this, we need to know the resulting value of $G \cdot \underline{\Delta}\,\underline{\mu}$, the aggregated individual-specific error terms.

From the above definitions,

$$G \cdot \underline{\Delta}\,\underline{\mu} = \left\{ I_T \otimes (I_m \otimes 1/n.\ 1_n')\ (1_T \otimes I_N) \right\}$$

$$= \left\{ I_T \cdot 1_T \otimes ((I_m \otimes 1/n.\ 1_n')\ I_N) \right\} \underline{\mu}$$

$$= \left\{ 1_T \otimes (I_m \otimes 1/n.\ 1_n') \right\} \underline{\mu}$$

$$= \begin{bmatrix} I_m \otimes 1/n.\ 1_n' \\ \vdots \\ I_m \otimes 1/n.\ 1_n' \end{bmatrix} \underline{\mu}$$

The last expression may be seen to be equivalent to $(1_T \otimes I_m)\ \bar{\mu}$, where $\bar{\mu}$ is the n x 1 vector containing the averages of the μ_i's corresponding to the individuals aggregated in a farticular group. Notice that $1_T \otimes I_m$ is of the same form as $\underline{\Delta}$, so that the aggregated model can be adequately represented with m "dummies" corresponding to each group, and "coefficients" μ_j, j = 1,..... m that consist of averages of the n individual error terms within a group. It follows that the methods for dealing with pooled models (LSDV or GLS) may be applied without change to data aggregated in this manner.

Table 1

Mean Errors of Estimate

for $Y = \beta_1 X_1 + \beta_2 X_2 + u$

(correct specification)

$R^2 = .7$

r_{12} =	.25		.50		.75	
	$\overline{b}_1$	$\overline{b}_2$	$\overline{b}_1$	$\overline{b}_2$	$\overline{b}_1$	$\overline{b}_2$
Ungrouped	-.004	.007	-.006	.009	-.010	.013
Random grouping	-.004	-.001	-.005	.001	-.008	.004
Grouping by X_1	.001	-.006	-.018	.035	.034	-.045
Grouping by X_2	.013	.002	-.002	.006	-.022	.022

$R^2 = .3$

r_{12} =	.25		.50		.75	
	$\overline{b}_1$	$\overline{b}_2$	$\overline{b}_1$	$\overline{b}_2$	$\overline{b}_1$	$\overline{b}_2$
Ungrouped	-.009	.017	-.013	.021	-.022	.030
Random grouping	-.010	-.001	-.012	.003	-.018	.010
Grouping by X_1	.001	-.015	-.043	.082	.079	-.105
Grouping by X_2	.031	.005	-.006	.014	-.052	.050

Table 2

Mean Errors of Estimate

for $Y = b_1X_1 + v$

(misspecified model)

$R^2 = .7$

r_{12} =	.25	.50	.75
	b_1	b_1	b_1
Ungrouped	.246	.498	.751
Random grouping	.249	.502	.755
Grouping by X_1	.246	.498	.751
Grouping by X_2	1.667	1.571	1.253

$R^2 = .3$

r_{12} =	.25	.50	.75
	b_1	b_1	b_1
Ungrouped	.243	.496	.751
Random grouping	.244	.498	.752
Grouping by X_1	.244	.497	.751
Grouping by X_2	1.691	1.583	1.261

Table 3

Proportion of Positively Biased Estimates

for $Y' = \beta_1 X_1 + \beta_2 X_2 + u$

(correct specification)

$R^2 = .7$

r_{12} =	.25		.50		.75	
	$\overline{b_1}$	$\overline{b_2}$	$\overline{b_1}$	$\overline{b_2}$	$\overline{b_1}$	$\overline{b_2}$
Ungrouped	.44	.56	.45	.59	.46	.56
Random grouping	.49	.48	.47	.48	.47	.49
Grouping by X_1	.52	.45	.41	.55	.60	.42
Grouping by X_2	.58	.48	.53	.57	.46	.55

$R^2 = .3$

r_{12} =	.25		.50		.75	
	$\overline{b_1}$	$\overline{b_2}$	$\overline{b_1}$	$\overline{b_2}$	$\overline{b_1}$	$\overline{b_2}$
Ungrouped	.44	.56	.45	.59	.46	.56
Random grouping	.49	.48	.47	.48	.47	.49
Grouping by X_1	.52	.45	.41	.55	.60	.42
Grouping by X_2	.58	.48	.53	.57	.46	.55

Table 4

Proportion of Positively Biased Estimates

for $Y = b_1X_1 + v$

(misspecified model)

$R^2 = .7$

r_{12} =	.25	.50	.75
	$\overline{b_1}$	$\overline{b_1}$	$\overline{b_1}$
Ungrouped	1.00	1.00	1.00
Random grouping	.90	1.00	1.00
Grouping by X_1	1.00	1.00	1.00
Grouping by X_2			

$R^2 = .3$

r_{12} =	.25	.50	.75
Ungrouped	.99	1.00	1.00
Random grouping	.70	.89	.96
Grouping by X_1	.99	1.00	1.00
Grouping by X_2	1.00	1.00	1.00

Table 5

Mean Squared Errors of Estimate

for $Y = \beta_1 X_1 + \beta_2 X_2 + u$

(correct specification)

$R^2 = .7$

r_{12} =	.25		.50		.75	
	$\overline{b_1}$	$\overline{b_2}$	$\overline{b_1}$	$\overline{b_2}$	$\overline{b_1}$	$\overline{b_2}$
Ungrouped	.002	.002	.003	.003	.007	.007
Random grouping	.027	.025	.043	.039	.090	.081
Grouping by X_1	.003	.019	.012	.037	.034	.057
Grouping by X_2	.027	.004	.041	.013	.078	.043

$R^2 = .3$

r_{12} =	.25		.50		.75	
	$\overline{b_1}$	$\overline{b_2}$	$\overline{b_1}$	$\overline{b_2}$	$\overline{b_1}$	$\overline{b_2}$
Ungrouped	.012	.011	.018	.018	.037	.038
Random grouping	.147	.135	.236	.210	.492	.443
Grouping by X_1	.016	.102	.063	.199	.196	.311
Grouping by X_2	.147	.019	.224	.071	.427	.234

Table 6

Mean Square Errors of Estimate

for $Y = b_1X_1 + v$

(misspecified model)

$R^2 = .7$

r_{12} =	.25	.50	.75
	b_1	b_1	b_1
Ungrouped	.064	.252	.568
Random grouping	.106	.295	.611
Grouping by X_1	.064	.252	.568
Grouping by X_2	2.842	2.484	1.577

$R^2 = .3$

r_{12} =	.25	.50	.75
	b_1	b_1	b_1
Ungrouped	.071	.260	.579
Random grouping	.203	.408	.746
Grouping by X_1	.072	.261	.579
Grouping by X_2	2.978	2.549	1.618

Table 7

Comparison of Alternative Specifications of a Reading Achievement Regression: Ungrouped and Grouped Data (beta-weights)

	Level of Analysis				
Independent Variables	Pupil	School	District	Pupil	School
Resources	-.119	-.465	-.664	-.149	-.196
Pupil/Teacher Ratio	-.167	-.368	-.737	-.146	-.048
% Non-White	-.292	-.007	-.085	-.109	-.198
Parents' Occupation (1)				.037	.143
Parents' Occupation (2)				.126	-.072
Parents' Occupation (3)				.160	.255
Parents' Occupation (4)				.202	.464
Parents' Occupation (5)				.243	.051
I.Q. Estimate				.385	.081

References

Amemiya, Takeshi
1967 "A note on the estimation of Balestra-Nerlove Models." Technical Report No. 4. Stanford University: Institute for Mathematical Studies in the Social Sciences.

Bidwell, Charles E. and John D. Kasarda
1975 "School district organization and student achievement." American Sociological Review 40 (February): 55-70.

Blalock, Hubert M.
1964 Causal Inferences in Nonexperimental Research. Chapel Hill, N. C.: Univeristy of North Carolina Press.

Cramer, J.S.
1964 "Efficient grouping: regression and correlation in Engel curve analysis." Journal of the American Statistical Association 59 (March): 233-50.

Feige, Edgar L. and Harold W. Watts
1972 "An investigation of the consequences of partial aggregation of micro-economic data." Econometrics 40 (March): 343-60.

Hannan, Michael T.
1971 Aggregation and Disaggregation in Sociology. Lexington, Mass.: Heath.

Hannan, Michael T. and Leigh Burstein
1974 "Estimation from grouped observations." American Journal of Sociology 39 (June): 374-92.

Hannan, Michael T. and Alice A. Young
1974 "Estimation in panel models: results on pooling cross-sections and time series." Stanford University: Laboratory for Social Research, Technical Report #51.

Hanushek, Eric A., John E. Jackson and John F. Kain
1974 "Model specification, use of aggregate data, and the ecological correlation fallacy." Political Methodology 1 (August): 89-107.

Johnston, J.
1972 Econometric Methods. 2nd edition, New York: McGraw-Hill.

Nerlove, Marc
1971 "Further evidence on the estimation of dynamic economic relations from a time series of cross-sections." Econometrics 39 (March): 359-382.

Prais, S.J. and J. Aitchison
1954 "The grouping of observations in regression analysis." Review of the International Statistical Institute 22: 1-22.

References cont'd.

Theil, Henri
1957 "Specification errors and the estimation of economic relationships." Review of the International Statistical Institute 25 (August): 41-51.